SketchFlow: Zero-Shot Vector Sketch Generation via GMM Prior Flow in CLIP Latent Space

ANONYMOUS AUTHOR(S)
SUBMISSION ID: 2082

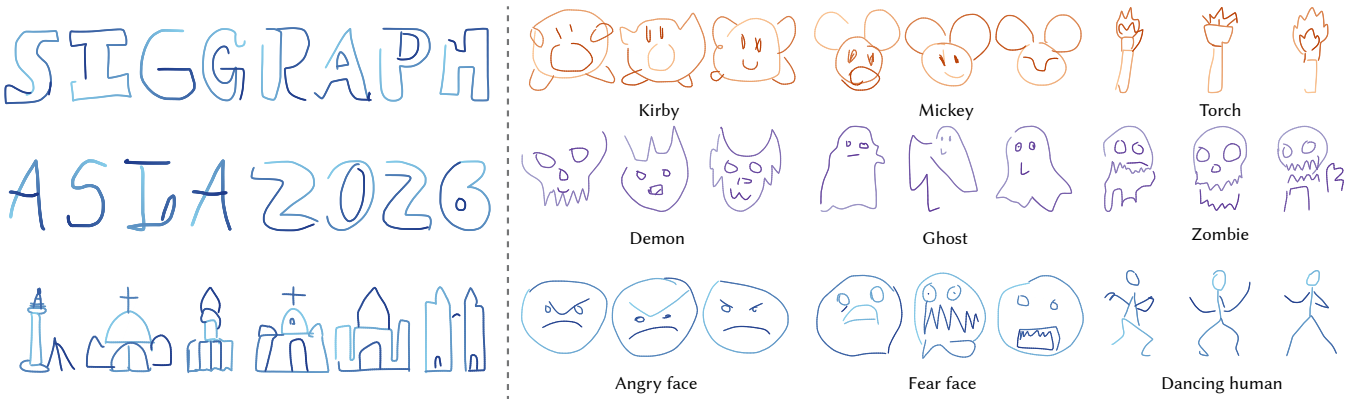

Fig. 1. **Open-vocabulary vector sketch generation by SketchFlow. SketchFlow** excavates drawing priors from the CLIP latent space for zero-shot synthesis. Left: Text-prompted generation of letters, numbers, and landmarks in Malaysia. Right: Diverse zero-shot generation including pop-culture characters, creatures, abstract emotions, and dynamic poses. **SketchFlow** successfully mimics human drawing behaviors without category-specific training data.

Vector sketches remain one of the most concise and immediate mediums for abstract human expression. However, generating high-quality vector strokes that exhibit human-like drawing styles remains an open challenge due to the severe scarcity of fine-grained, high-quality text-to-sketch paired data. Existing text-conditioned generation methods often rely on unstable, time-consuming optimization or struggle to generalize to unseen categories in a zero-shot manner. To address these limitations, we present **SketchFlow**, a novel generative framework rooted in Optimal Transport (OT) theory and flow matching. By leveraging pre-trained CLIP models to bypass labor-intensive image-level text annotations, we formulate cross-modal alignment as a continuous mapping problem directly within the CLIP latent space. To bridge the inevitable modality gap between discrete text concepts and continuous sketch features, we first inject noise into discrete category embeddings to construct a continuous Gaussian Mixture Model (GMM) prior. We then utilize an Optimal Transport Conditional Flow Matching (OT-CFM) model to learn a deterministic vector field mapping from this continuous GMM prior to the target sketch feature distribution. Finally, a Hybrid Diffusion Decoder, fusing 1D U-Net and Transformer architectures, is designed to decode these features into fast and high-fidelity stroke trajectories. Extensive experiments demonstrate that **SketchFlow** substantially outperforms existing baselines in visual quality and adherence to natural human drawing styles. Furthermore, our geometry-preserving framework exhibits robust zero-shot Open-vocabulary synthesis capabilities, effectively responding to out-of-domain prompts, incorporating diverse semantic modifiers, and enabling smooth, continuous semantic interpolation between distinct concepts.

ACM 1557-7368/2026/5-ART
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Anonymous Author(s). 2026. **SketchFlow**: Zero-Shot Vector Sketch Generation via GMM Prior Flow in CLIP Latent Space. *ACM Trans. Graph.* 1, 1 (May 2026), 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Vector sketches are among the most concise and direct mediums for abstract human expression. Consequently, text-conditioned vector sketch synthesis holds significant potential in computer graphics domains, including digital avatar creation, conceptual design, and interactive art. However, generating high-quality vector strokes that exhibit human-like drawing styles remains an open challenge. Existing mainstream methods suffer from non-negligible limitations. For optimization-based approaches [Arar et al. 2025; Polaczek et al. 2025; Xing et al. 2023], they are typically time-consuming, unstable, and prone to producing noisy results. Meanwhile, raster-based generation methods [Hu et al. 2024] inherently lack stroke topological constraints. They often produce unvectorizable artifacts such as inconsistent thickness, semi-transparency, or unintended branching, which renders complex post-processing steps (e.g., vector tracing) frequently ineffective. Moreover, recent direct sequence generation methods based on Large Language Models (LLMs) [Vinker et al. 2025], while avoiding image rendering, produce stroke trajectories that deviate significantly from authentic human drawing habits in terms of style and abstract logic.

The most intuitive approach to this task is end-to-end training using large-scale, paired text-vector sketch data. However, high-quality annotated data of this nature is exceedingly scarce. Although QuickDraw [Ha and Eck 2017], currently the largest vector sketch dataset, contains a massive collection of human-drawn trajectories, it provides only 345 discrete category labels and lacks fine-grained textual descriptions.

Consequently, we confront a fundamental research gap: How can we achieve zero-shot, text-to-sketch synthesis for open-world concepts using only training sketches with discrete category labels? To address this, we present **SketchFlow**, a novel generative framework rooted in Optimal Transport theory and flow matching, which constructs continuous, deterministic trajectories between distributions. By formulating cross-modal alignment as a continuous mapping problem, **SketchFlow** enables zero-shot vector sketch generation via a GMM prior flow embedded directly in the CLIP latent space.

Our core insight is that by leveraging pre-trained CLIP models, we can bypass labor-intensive image-level text annotation, enabling direct semantic alignment in a high-level conceptual space. We formulate this task as a cross-modal latent space mapping problem. Due to the inevitable modality gap between text and sketches in the CLIP embedding space, directly equating discrete text features with continuous sketch visual features often leads to generation failure (see Figure 7). To bridge this gap, we first inject noise into the discrete category embeddings to construct a continuous Gaussian Mixture Model (GMM) prior distribution. This crucial step expands isolated semantic points into a continuous probability space, providing a smooth base distribution. Subsequently, we introduce an Optimal Transport Conditional Flow Matching (OT-CFM) model to learn the optimal vector field mapping from this continuous GMM prior to the target sketch feature distribution. Finally, to decode these features into concrete stroke sequences, we design a hybrid diffusion decoder fusing a 1D sequence-adapted U-Net and Transformer architectures, achieving fast and high-fidelity stroke trajectory reconstruction.

Extensive experimental results demonstrate the superiority of our approach. This work provides a practical framework for vector sketch generation and highlights the potential of flow matching techniques in cross-modal zero-shot generation. Our core contributions are summarized as follows:

- **Zero-shot Open-vocabulary synthesis:** Our model exhibits remarkable zero-shot capabilities for out-of-domain textual prompts, accurately responding to complex descriptions that differ significantly from training categories, and supporting additional modifiers to guide the synthesis of sketches with specific attributes.
- **Novel cross-modal mapping framework:** We introduce an Optimal Transport Conditional Flow Matching model combined with a continuous GMM prior to effectively bridge the modality gap between discrete text concepts and continuous sketch features in the CLIP latent space.
- **High-fidelity stroke reconstruction:** We design a hybrid diffusion decoder fusing 1D U-Net and Transformer architectures, which outperforms existing baselines in visual quality and adherence to human drawing styles.

2 Related Work

2.1 Vector Sketch Generation

Vector sketches possess an inherent temporal structure; thus, early research naturally formulated the generation process as trajectory prediction over ordered strokes. SketchRNN [Ha and Eck 2017] pioneered this approach by introducing the QuickDraw dataset alongside an LSTM-VAE architecture for sketch encoding and decoding. Subsequent works [Aksan et al. 2020; Ashcroft et al. 2023; Bandyopadhyay et al. 2024; Das et al. 2021, 2023; Wang et al. 2023; Xu et al. 2024] have refined stroke modeling by utilizing RNNs, Ordinary Differential Equations (ODEs), and sequence diffusion. While these models effectively capture temporal dynamics, they typically support only unconditional or few-shot conditional generation, which severely limits their scalability and text controllability. Although Scale-Adaptive Diffusion [Hu et al. 2024] demonstrated large-scale training capabilities on the QuickDraw dataset, it generates raster images rather than scalable vector sketches. More recently, StrokeFusion [Zhou et al. 2026] proposed training a diffusion model to denoise and generate stroke sets; however, it requires a manually specified upper bound for the number of strokes and struggles to generalize to unseen categories in a zero-shot manner.

2.2 Text-Conditioned Visual Synthesis

Recent controllable generation methods integrate pre-trained vision-language models to impose semantic guidance. One prominent line of work performs direct text alignment via CLIP. Building upon differentiable vector rendering [Li et al. 2020], methods such as CLIPDraw [Frans et al. 2022], CLIPasso [Vinker et al. 2022], and CLIPascene [Vinker et al. 2023] optimize Bézier curves to maximize text-image similarity. Since these methods operate without ground-truth data references, they fail to produce authentic hand-drawn style, and direct CLIP similarity optimization is notoriously unstable.

Another branch leverages Score Distillation Sampling (SDS) [Poole et al. 2022] to achieve text-guided optimization in the vector domain. DiffSketcher [Xing et al. 2023] utilizes latent diffusion and SDS for unpaired text-to-vector synthesis, while SwiftSketch [Arar et al. 2025] introduces an efficient diffusion architecture incorporating stroke-order constraints. Furthermore, style customization based on diffusion priors [Zhang et al. 2025] and implicit curve optimization via NeuralSVG [Polaczek et al. 2025] have been explored. Despite their visual fidelity, SDS-based methods fundamentally rely on heavy, time-consuming inference-time optimization. In contrast, our approach operates as a highly efficient forward-pass generation framework.

Alternative methods include SketchAgent [Vinker et al. 2025], which explores LLMs for generating stroke sequences. Although LLMs maintain logical drawing orders, their outputs struggle to mimic natural brushstrokes, often resembling rigid SVG graphics. Moreover, while works like IconShop [Wu et al. 2023] have explored SVG generation using paired data, acquiring large-scale, cleanly annotated paired data for highly abstract sketches remains highly challenging.

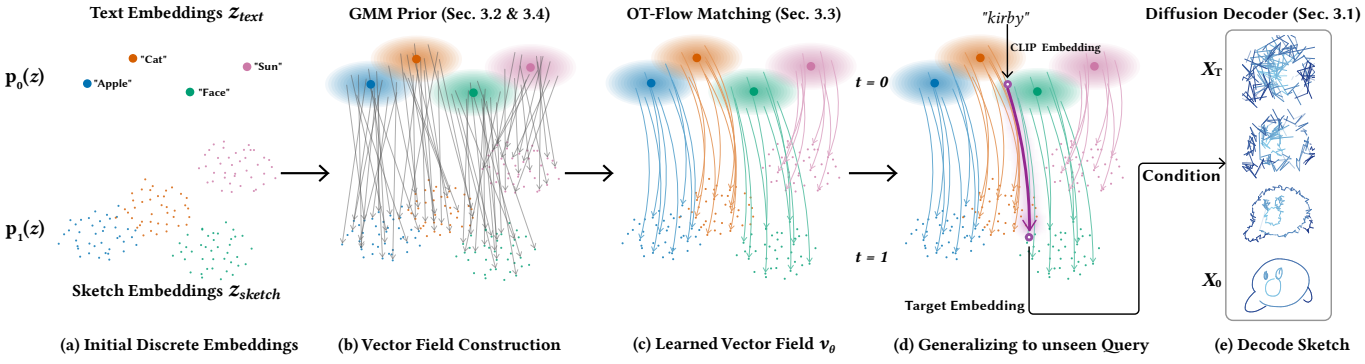

Fig. 2. **Overview of the SketchFlow generative framework.** (a) **Initial Discrete Embeddings:** The framework extracts discrete text embeddings z^k_{text} and sketch embeddings for known categories in the latent space. (b) **Vector Field Construction:** A noise injection mechanism is introduced to transform sparse concept anchors into a continuous probability density field, establishing a Gaussian Mixture Model (GMM) prior (Sec. 3.2 & 3.4). (c) **Learned Vector Field v_θ:** Optimal Transport Conditional Flow Matching establishes a deterministic mapping between the GMM prior $p_0(z)$ and the target sketch distribution, naturally yielding straight-path trajectories (Sec. 3.3). (d) **Generalizing to Unseen Query:** During inference, the CLIP embedding of an unseen query (e.g., "kirby") smoothly follows the transport trajectories of adjacent known concepts to safely land on the valid sketch manifold. (e) **Decode Sketch:** Conditioned on the acquired target sketch embedding, a Hybrid Diffusion Decoder utilizes a reverse diffusion process to decode pure noise X_T into a final high-fidelity vector stroke sequence X_0 (Sec. 3.1).

2.3 Flow Matching and Optimal Transport

Flow Matching [Lipman et al. 2022] has recently emerged as a powerful alternative to diffusion models. By introducing the Conditional Flow Matching (CFM) framework, it has demonstrated superior performance and stability in deterministic generation tasks. Recent advancements have deeply explored its Optimal Transport (OT) properties; methods such as Rectified Flow [Liu et al. 2022] and Optimal Transport Conditional Flow Matching (OT-CFM) [Tong et al. 2023] significantly enhance sampling efficiency by linearizing probability paths.

Recent advancements have addressed the limitations of standard Gaussian priors by employing Gaussian Mixture Models (GMMs) as Conditional Prior Distributions (CPD) [Issachar et al. 2025]. By minimizing the optimal transport cost, GMM priors substantially improve the modeling of complex data distributions. Building upon these insights, we pioneer the application of GMM-based flow matching to zero-shot vector sketch generation. Specifically, we embed a continuous GMM prior within the CLIP latent space. This approach effectively translates isolated, discrete textual categories into a continuous semantic distribution, thereby bridging the cross-modal gap between textual concepts and continuous visual trajectories.

3 Method

3.1 Sketch Representation and Hybrid Diffusion Decoder

To better understand the required inputs for sketch synthesis, we first introduce the final generation module: the Hybrid Diffusion Decoder. As illustrated in the overall framework (Figure 2), to decode target CLIP embeddings into vector stroke sequences, we design a conditional feature decoder. The sketch trajectory is formulated as an equidistant sequence of L points, denoted as $S = \{(x_i, y_i, p_i)\}_{i=1}^{L}$. Here, (x_i, y_i) represents the 2D spatial coordinates. To facilitate a continuous diffusion process and avoid early over-commitment to discrete states at high-noise stages, we treat the pen-state indicator as a continuous variable $p_i \in \mathbb{R}$, scaled to a small variance range (e.g., $\{-0.1, 0.1\}$) rather than strictly binary. This ensures that p_i specifies connectivity to the preceding point without interfering with the synthesis of the global geometric trajectory during the early generation steps.

For the generative backbone, we employ a Hybrid Diffusion Decoder (Figure 3) featuring an interleaved design of 1D U-Net convolutional blocks [Ronneberger et al. 2015] and Transformer layers [Vaswani et al. 2017]. Within this architecture, the U-Net-style operations handle local geometric feature extraction and multi-resolution downsampling/upsampling, while the interleaved Transformer blocks process long-range global dependencies at their respective resolution scales.

For the conditioning mechanism, the diffusion timestep t and the target sketch CLIP embedding (denoted as z_{target}) are independently processed via sinusoidal positional encodings and linear projections. These representations are then summed to form a unified condition vector, which is injected into the hidden features of each U-Net and Transformer block using Adaptive Layer Normalization (AdaLN). The decoder is trained using a standard denoising score matching objective [Ho et al. 2020; Song et al. 2020]:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(x_t, t, z_{\text{target}})\right\|^2\right]$$

We observe that when conditioned on a valid sketch CLIP embedding, this decoder inherently exhibits robust zero-shot generalization, capable of generating coherent strokes for unseen concepts. Therefore, the primary objective of our framework becomes acquiring the target sketch CLIP embedding directly from a textual prompt. We address this modality gap by constructing a continuous prior in the CLIP latent space, as detailed in the following section.

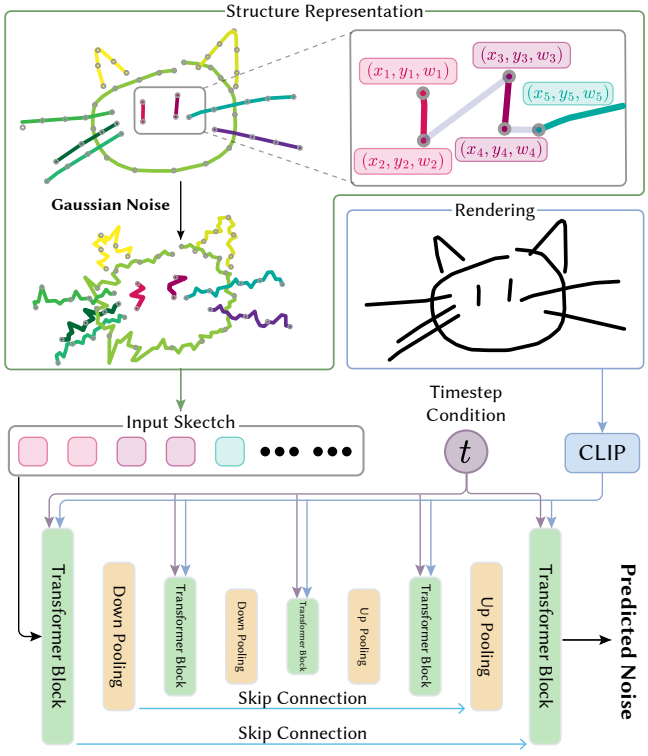

Fig. 3. **Architecture of the Hybrid Diffusion Decoder.** The model employs an interleaved generative backbone combining U-Net-style bottlenecks for local geometric extraction with Transformer blocks for global dependency modeling. Condition embeddings (diffusion timestep t and target CLIP embedding) are injected via AdaLN to predict the added noise.

3.2 GMM Prior Construction in CLIP Latent Space

To bypass the scarcity of high-quality, fine-grained paired text-sketch data, we formulate the cross-modal alignment as a distribution transport problem in the CLIP latent space. Given that available datasets predominantly provide categorical labels rather than rich textual descriptions, we leverage the pre-trained CLIP text encoder [Radford et al. 2021] to extract discrete text embeddings z^k_{text} for each category k.

However, treating these isolated text embeddings as direct initial states fails to cover the continuous semantic space necessary for open-world generation. To bridge this modality gap, we introduce a noise injection mechanism to construct a continuous Gaussian Mixture Model (GMM) prior. At initial state $t = 0$, we define the prior distribution $p_0(z)$ as:

$$p_0(z) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(z; z^k_{\text{text}}, \sigma^2 I)$$

This formulation transforms sparse concept anchors into a continuous probability density field, establishing a smooth semantic manifold from which our transport mapping originates. Here, σ represents the standard deviation of the injected noise. We treat σ as an adaptive parameter to dynamically modulate the prior, as elaborated in Sec. 3.4.

3.3 Cross-Modal Alignment via OT-Flow Matching

Given the constructed GMM prior $p_0(z)$ and the target sketch empirical distribution $p_1(z)$, our goal is to establish a deterministic mapping between these two modalities.

Existing conditional paradigms (e.g., unCLIP [Ramesh et al. 2022]) typically assume a standard normal prior $p_0(z) = \mathcal{N}(0, I)$ and treat the text embedding as a condition. However, when trained on data limited to discrete categorical clusters, such formulations struggle to generalize. Ambiguous queries often cause the model to output a probabilistic mixture of isolated distributions, meaning it stochastically snaps to known modes rather than achieving true semantic interpolation.

To address this limitation, we directly formulate the continuous GMM as the base prior $p_0(z)$ and utilize Optimal Transport (OT) Conditional Flow Matching [Lipman et al. 2022; Tong et al. 2023]. The generative process is governed by the ordinary differential equation (ODE) $\mathrm{d}z_t = v_t(z_t)\mathrm{d}t$. Crucially, rather than learning an arbitrary flow, we leverage an OT coupling between p_0 and p_1 to construct the target vector field. In practice, to maintain semantic consistency during training, we construct the coupling conditionally based on semantic categories. Specifically, given a category k, we independently sample the initial state $z_0 \sim \mathcal{N}(z^k_{\text{text}}, \sigma^2 I)$ (σ is the adaptive prior variance detailed in Sec. 3.4) and the target state $z_1 \sim p_1(z|k)$. The flow matching model is then trained to regress the optimal transport vector field, which trivially forms a straight path $z_0 \to z_1$, using the Conditional Flow Matching objective [Lipman et al. 2022]:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, p_0(z_0|k), p_1(z_1|k)} \left[\| v_\theta(z_t, t, c, \sigma) - (z_1 - z_0) \|^2 \right]$$

where the intermediate state is parameterized as $z_t = (1-t)z_0 + tz_1$.

This OT formulation minimizes the transport cost, naturally yielding straight-path, constant-velocity trajectories. This geometric constraint provides two highly desirable properties: (1) **Semantic Direction Preservation:** The straight-path constraint ensures that the inherent semantic feature directions within the CLIP latent space are well-maintained throughout the mapping process. (2) **Natural Zero-Shot Generalization:** By smoothly following the established transport trajectories of known concepts, the model naturally generalizes to unseen open-world queries.

3.4 Adaptive Prior Variance Injection

Empirical results suggest that a single fixed noise scale σ is suboptimal for constructing the initial GMM prior across all semantic categories. Different semantic concepts intrinsically require varying degrees of spatial dispersion to achieve optimal generation fidelity.

To address this, we introduce a tunable prior variance mechanism. Because our core objective is to establish a continuous mapping between two distinct distributions, we explicitly parameterize the conditional vector field of the flow matching model as $v_\theta(z_t, t, c, \sigma)$. Specifically, to construct this conditioning mechanism, the flow timestep t, the text embedding c, and the logarithm of the prior variance $\log(\sigma)$ are independently processed via sinusoidal positional encodings and linear projections. These representations are then

summed to form a unified condition vector, which is injected into the hidden features of the flow matching network using Feature-wise Linear Modulation (FiLM) [Perez et al. 2018].

This allows the generated marginal distribution at the endpoint to be dynamically modulated, denoted as $\hat{p}_1(z \mid z_{\text{test}}, \sigma)$. During training, we randomly sample standard deviations σ to modulate the initial GMM prior, directly shaping the generative trajectory. During inference, this formulation allows users to adaptively adjust σ based on specific prompts. Specifically, we introduce a scaling parameter γ to control the sampling standard deviation of the query point relative to the standard deviation of the input condition.

4 Experiments

4.1 Experimental Setup

Datasets. We train our model on the complete QuickDraw dataset, encompassing 345 categories and approximately 50 million sketches. We extract sequential stroke trajectories from the dataset for sequence-level modeling. Furthermore, we conducted experiments on cross-domain transfer datasets to evaluate the model's generalizability; these extensive results are provided in the supplementary material.

Implementation Details. For data preprocessing, we resample sketch sequences to $L = 256$ and scale binary pen-state indicators to $\{-0.1, 0.1\}$ to prevent premature commitment to discrete states during high-noise diffusion stages. During training, sketches are rendered in black and white for CLIP feature extraction via the pre-trained ViT-B-32. For visual presentation in this paper, generated stroke trajectories are explicitly color-coded from light to dark to illustrate the sequential drawing order. The OT-Flow Matching model is parameterized by a residual MLP with FiLM layers, while the Hybrid Decoder utilizes a 1D U-Net interleaved with Transformer blocks. Comprehensive architectural and training hyperparameters are provided in the supplementary material.

Evaluation Metrics. Visual fidelity is evaluated using FID on 256×256 rasterized renderings. Cross-modal semantic alignment is measured by CLIP Score using the prompt "a sketch of [category]". To assess trajectory abstraction, we evaluate the *RDP Point Count* by computing the absolute number of remaining anchor points after applying the Ramer-Douglas-Peucker algorithm [Ramer 1972] with a fixed tolerance. Fewer remaining points indicate higher abstraction, avoiding biases from the original sketch resolution.

Baselines. We compare our framework against representative methods across different generative paradigms: (1) **SketchRNN** [Ha and Eck 2017], the pioneering LSTM-VAE trajectory baseline; (2) **StrokeFusion** [Zhou et al. 2026], the state-of-the-art latent diffusion model for trajectories; (3) **NeuralSVG** [Polaczek et al. 2025], an SDS-based optimization approach; and (4) **SketchAgent** [Vinker et al. 2025], an LLM-driven interactive planning agent. To ensure fair comparison, we omit FID scores for optimization-based and LLM-driven methods, as they are not explicitly trained on our data distribution.

4.2 Comparison with Existing Methods

To comprehensively evaluate our approach against existing state-of-the-art methods, we select four representative categories: *Bus*, *Cat*, *Chair*, and *Train*. These categories exhibit relatively complex structural characteristics and are included in the known classes of the QuickDraw dataset, facilitating a fair and direct comparison with other QuickDraw-based methods.

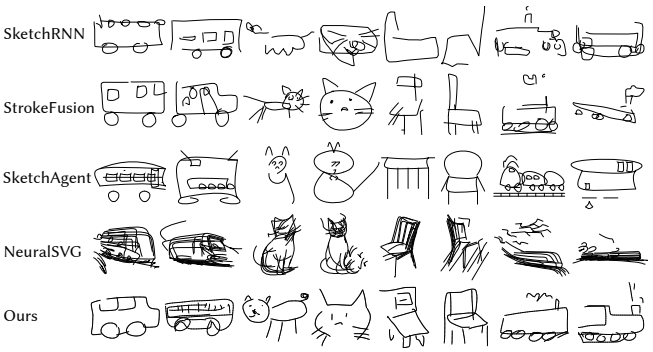

Fig. 4. **Qualitative comparison on QuickDraw.** Generated samples for four categories (*Bus*, *Cat*, *Chair*, and *Train*) from baseline methods and ours.

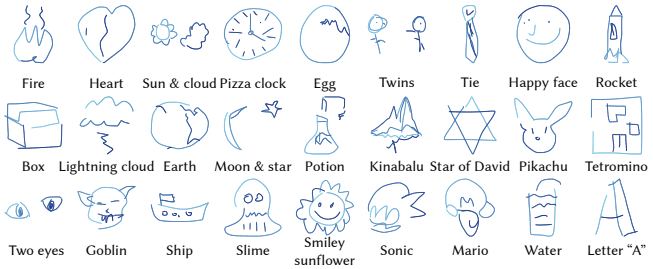

Fig. 5. **Open-vocabulary generation results from SketchFlow.** All prompts shown here lie outside the QuickDraw taxonomy, yet the model produces coherent and style-consistent vector sketches.

For each category, we generate 10,000 samples and compute the three evaluation metrics introduced in Sec. 4. To generate a specific category, the category name is used as the conditioning text input. Additional stochasticity is introduced by adding Gaussian noise scaled by the factor γ, as defined in Sec. 3.4. We report our results under varying noise scales $\gamma \in \{0.6, 0.8, 1.0\}$, while the base prior variance σ is fixed at 0.025. The corresponding quantitative evaluations and qualitative visualizations are presented in Table 1 and Figure 4, respectively.

As shown in Table 1, our method achieves significant improvements in the FID metric compared to all learning-based baselines across all categories. This substantial margin demonstrates that our model possesses superior representation and feature learning capabilities. Although some baseline methods achieve lower RDP scores, this is primarily because they tend to learn and generate overly simplistic or rudimentary sketches, which is demonstrated by the qualitative results in Figure 4.

Furthermore, our generated results achieve CLIP scores comparable to those of NeuralSVG and SketchAgent. However, the higher RDP scores of these baseline methods suggest a larger number of points after simplification, which often corresponds to cluttered

Fig. 6. **Zero-Shot Adjective Comprehension.** Generation results for specific action prompts. Three distinct results are shown for each prompt.

and unstructured strokes. While such visual complexity may inadvertently inflate CLIP recognition scores, it deviates from actual human drawing behaviors. In contrast, qualitative comparisons in Figure 4 illustrate that **SketchFlow** better captures the abstract characteristics of hand-drawn sketches, producing trajectories that are concise, coherent, and effective at conveying target concepts.

4.3 Zero-Shot and Open-vocabulary Generation

Beyond generating known classes from the training dataset, our model demonstrates a strong ability to generalize to unseen categories via OT-Flow Matching. Figure 5 presents a diverse set of open-vocabulary generation results, with additional examples provided in the supplementary material. The model synthesizes not only concepts that share structural similarities with QuickDraw categories but also those that differ significantly, such as *Pikachu*, *Tetromino*, and *Goblin*. Furthermore, it accurately interprets composite and plural concepts, as seen in the generation results for *Sun & cloud* and *Two eyes*. This demonstrates that our framework can effectively excavate the visual drawing priors of these complex concepts directly from the CLIP latent space.

Furthermore, our model exhibits a robust understanding of semantic modifiers and actions, even though it was trained exclusively on discrete category names. As demonstrated in Figure 6, when conditioned on prompts with specific descriptive adjectives or actions (e.g., *a running cat*, *a jumping cat*, *a stretching cat*), the model generates trajectories that accurately reflect these dynamic poses while maintaining the abstract hand-drawn style. This highlights the powerful zero-shot semantic alignment enabled by our continuous GMM prior construction and flow-matching mapping.

4.4 User Studies

To evaluate the perceptual quality and human-likeness of **SketchFlow**, we conducted a user study following [Bhunia et al. 2022; Wang et al. 2024], comparing our approach against two representative baselines: NeuralSVG [Polaczek et al. 2025] and SketchAgent [Vinker et al. 2025]. To account for the variance in generation quality among these methods and ensure a fair evaluation, we selected the single best result (highest CLIP score) from 10 generated samples per prompt. A total of 42 participants were recruited for this study. They assessed the sketches across five metrics: abstraction fidelity, semantic alignment, human-likeness, stroke rationality, and aesthetic quality. As shown in Table 2, **SketchFlow** consistently outperforms the comparison methods across all criteria. This strong preference demonstrates that our approach successfully captures the abstract semantic essence of concepts and naturally simulates authentic human drawing nuances (*e.g.*, slight stroke variations).

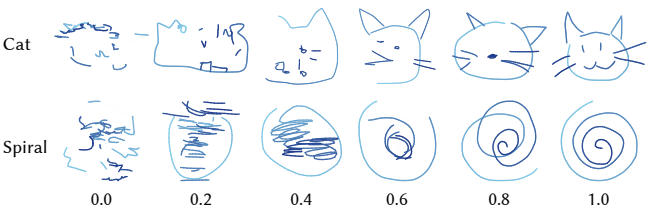

Fig. 7. **Ablation of Flow Mapping.** Results of linear interpolating the condition from raw text embeddings ($\lambda = 0.0$) to Flow-mapped embeddings ($\lambda = 1.0$) at intervals of 0.2. The raw text embedding alone is insufficient to guide structured sketch synthesis.

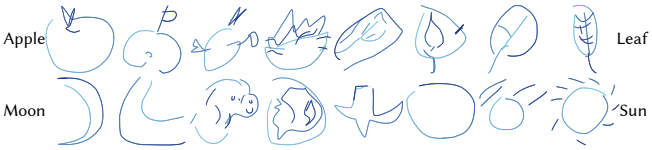

Fig. 8. **Ablation of Gaussian-Prior baseline** Latent interpolation results produced by the baseline. Unlike our approach, this baseline struggles to maintain structural coherence during the transition. The intermediate sketches exhibit abrupt topological shifts and a loss of visual semantics, highlighting the necessity of our GMM-based OT-Flow Matching design for establishing a continuous generative space.

4.5 Continuous Interpolation

To demonstrate that our framework learns a continuous semantic manifold rather than merely overfitting to discrete training categories, we perform linear interpolation between distinct text embeddings. As illustrated in Figure 9, the model generates smooth and topologically coherent structural transitions between entirely different concepts, validating the robust continuity of the established generative space.

4.6 Ablation Studies

A fundamental challenge in cross-modal generation is the inherent modality gap; the distribution of CLIP text embeddings does not naturally align with the distribution of CLIP sketch embeddings. Consequently, our proposed OT-Flow Matching module is an essential component designed to bridge this semantic divergence.

To explicitly demonstrate this necessity, we conduct an ablation study on the flow mapping process, as shown in Figure 7. We perform a linear interpolation in the latent space between the raw CLIP text embedding ($\lambda = 0.0$) and the corresponding target embedding mapped by our flow model ($\lambda = 1.0$). When the diffusion decoder is directly conditioned on the raw text embedding without flow mapping, it fails to land on the valid sketch manifold, producing unstructured and unrecognizable strokes. However, as the conditioning vector progressively shifts toward the flow-mapped output ($\lambda \to 1.0$), the generated trajectories smoothly self-organize into structurally coherent and recognizable sketches.

To empirically validate the necessity of our GMM prior formulation over the conventional conditional paradigm (Sec. 3.3), we train a variant denoted as the "Gaussian-Prior Baseline." This model

Method	*Bus*			*Cat*			*Chair*			*Train*		
	FID ↓	CLIP ↑	RDP ↓	FID ↓	CLIP ↑	RDP ↓	FID ↓	CLIP ↑	RDP ↓	FID ↓	CLIP ↑	RDP ↓
SketchRNN	66.103	0.266	**79.571**	70.704	0.210	**68.591**	88.294	0.227	25.351	83.578	0.236	**78.830**
StrokeFusion	59.329	0.278	86.690	52.940	0.260	85.730	33.463	0.277	36.460	59.418	0.247	102.910
SketchAgent	–	0.258	124.429	–	0.214	93.684	–	0.267	40.200	–	0.234	98.400
NeuralSVG	–	0.224	278.446	–	0.261	267.091	–	**0.304**	225.663	–	0.225	256.168
Gaussian-Prior	9.556	0.265	98.53	24.568	0.148	97.15	17.572	0.165	38.45	15.180	0.183	99.06
Ours ($\gamma = 0.60$)	19.381	**0.287**	85.47	27.977	**0.262**	79.77	41.123	0.294	**24.30**	15.347	**0.260**	92.29
Ours ($\gamma = 0.80$)	9.004	0.284	92.12	17.899	0.261	86.33	21.545	0.287	29.75	8.217	0.256	96.21
Ours ($\gamma = 1.00$)	**5.810**	0.276	100.35	**14.423**	0.254	95.75	**11.718**	0.280	36.65	**7.181**	0.248	103.79
Dataset	0	0.284	91.88	0	0.259	77.83	0	0.283	29.29	0	0.266	101.89

Table 1. Quantitative results on QuickDraw *Bus*, *Cat*, *Chair*, and *Train*. Bold and underlined values indicate the best and second-best performances, respectively.

Table 2. User study results comparing **SketchFlow** with **NeuralSVG** and **SketchAgent** across five metrics. The scores indicate the average percentage of user preference for each method.

Metric	**SketchFlow**	**NeuralSVG**	**SketchAgent**
Abstraction Fidelity	72.85%	12.52%	14.63%
Semantic Alignment	56.75%	32.52%	10.73%
Human-likeness	70.73%	16.59%	12.68%
Stroke Rationality	61.79%	26.18%	12.03%
Aesthetic Quality	51.71%	37.24%	11.05%

follows the unCLIP approach by assuming a standard normal prior $p_0(z) = \mathcal{N}(0, I)$ and treating the text embedding strictly as a condition.

The quantitative evaluation of this baseline is recorded in Table 1, where it exhibits a clear performance degradation compared to our method. This decline stems from the unCLIP paradigm's inability to explicitly align or organize the cross-modal distributions. Consequently, a perturbed conditional embedding from one category can easily drift into the latent representation of another, resulting in a noticeable probability of generating sketches that deviate from the input class.

More importantly, as discussed in Section 3.3, this purely conditional model fundamentally struggles to generalize to unseen categories. To illustrate this, we replicate the latent interpolation experiment using the unCLIP baseline, with results shown in Figure 8 and Figure 9. The generated sketches exhibit abrupt structural shifts and rapidly lose visual readability. Rather than traversing a continuous semantic manifold, the model simply snaps between isolated training modes, failing to synthesize structurally coherent intermediate concepts.

4.7 Image-Conditioned Generation

Exploiting the inherently aligned vision-language latent space of the CLIP model, our text-to-sketch framework seamlessly extends to image-conditioned generation without any architectural modifications or paired image-sketch fine-tuning. By substituting the input textual prompt with the CLIP visual embedding of a reference image, SketchFlow directly synthesizes vector strokes that abstract the visual input. As demonstrated in Figure 10, this formulation effectively functions as a zero-shot semantic abstraction and style transfer mechanism. The model bypasses complex prompt engineering, extracting not only the core semantic identity (e.g., a cat or a bus) but also adapting to the geometric structure, pose, and stylistic abstraction level of the reference image.

5 Conclusion & Limitations

In this paper, we presented **SketchFlow**, a novel generative framework designed to tackle the challenging problem of zero-shot text-to-vector sketch synthesis. By introducing an Optimal Transport (OT) Conditional Flow Matching model and constructing a continuous Gaussian Mixture Model (GMM) prior within the CLIP latent space, our approach effectively bridges the modality gap between discrete text concepts and continuous sketch features. This method successfully extends the zero-shot generation capabilities of models traditionally trained on discrete categorical datasets, demonstrating a substantial advancement in both generation quality and stylistic adherence to human drawing behaviors.

Crucially, our findings demonstrate that the geometric properties of Optimal Transport can be explicitly leveraged to train highly capable zero-shot models, pointing toward a future where generative systems might eventually significantly reduce their reliance on exhaustive, fine-grained textual annotations.

However, we acknowledge that our current framework still possesses limitations. The mapping from text to sketch remains imperfect; for certain unseen categories, the model requires an expanded search variance (γ) to locate appropriate generation trajectories, and the quality of the resulting syntheses can be inconsistent. We hypothesize that this limitation stems from our foundational assumption: the current framework implicitly assumes that the semantic flow near each categorical text anchor forms a relatively uniform Gaussian distribution. In reality, the true semantic manifold of different concepts is highly complex, individualized, and non-Gaussian, a nuance that our current GMM prior construction struggles to fully capture. Future research will focus on developing more sophisticated and precise latent mapping techniques to address these topological discrepancies.

References

Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, and Otmar Hilliges. 2020. Cose: Compositional stroke embeddings. *Advances in Neural Information Processing Systems* 33 (2020), 10041–10052.

Ellie Arar, Yarden Frenkel, Daniel Cohen-Or, Ariel Shamir, and Yael Vinker. 2025. Swiftsketch: A diffusion model for image-to-vector sketch generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–12.

Alexander Ashcroft, Ayan Das, Yulia Gryaditskaya, Zhiyu Qu, and Yi-Zhe Song. 2023. Modelling complex vector drawings with stroke-clouds. In *The Twelfth International Conference on Learning Representations*.

Hmrishav Bandyopadhyay, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Tao Xiang, Timothy Hospedales, and Yi-Zhe Song. 2024. Sketchinr: A first look into sketches as implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12565–12574.

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Jorma Laaksonen, and Michael Felsberg. 2022. Doodleformer: Creative sketch drawing with transformers. In *European Conference on Computer Vision*. Springer, 338–355.

Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. 2021. Sketchode: Learning neural sketch representation in continuous time. In *International Conference on Learning Representations*.

Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. 2023. Chirodiff: Modelling chirographic data with diffusion models. *arXiv preprint arXiv:2304.03785* (2023).

Kevin Frans, Lisa Soros, and Olaf Witkowski. 2022. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems* 35 (2022), 5207–5218.

David Ha and Douglas Eck. 2017. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477* (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Jijin Hu, Ke Li, Yonggang Qi, and Yi-Zhe Song. 2024. Scale-adaptive diffusion model for complex sketch synthesis. In *The Twelfth International Conference on Learning Representations*.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. 2025. Designing a Conditional Prior Distribution for Flow-Based Generative Models. *arXiv preprint arXiv:2502.09611* (2025).

Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–15.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Sagi Polaczek, Yuval Alaluf, Elad Richardson, Yael Vinker, and Daniel Cohen-Or. 2025. Neuralsvg: An implicit representation for text-to-vector generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15458–15468.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. DreamFusion: Text-to-3D using 2D Diffusion. In *The Eleventh International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

Urs Ramer. 1972. An iterative procedure for the polygonal approximation of plane curves. *Computer Graphics and Image Processing* 1, 3 (1972), 244–256.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. 2023. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482* (2023).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, and Ariel Shamir. 2023. Clipascene: Scene sketching with different types and levels of abstraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4146–4156.

Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. 2022. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–11.

Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. 2025. Sketchagent: Language-driven sequential sketch generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23355–23368.

Jiawei Wang, Zhiming Cui, and Changjian Li. 2024. VQ-SGen: A Vector Quantized Stroke Representation for Sketch Generation. *arXiv e-prints* (2024), arXiv–2411.

Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, and Yi-Zhe Song. 2023. Sketchknitter: Vectorized sketch generation with diffusion models. In *The eleventh international conference on learning representations*.

Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. 2023. Iconshop: Text-guided vector icon synthesis with autoregressive transformers. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–14.

Ximing Xing, Chuang Wang, Haitao Zhou, Jing Zhang, Qian Yu, and Dong Xu. 2023. Diffsketcher: Text guided vector sketch synthesis through latent diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 15869–15889.

Pengfei Xu, Banhuai Ruan, Youyi Zheng, and Hui Huang. 2024. Sketchformer++: A hierarchical transformer architecture for vector sketch representation. In *International Conference on Computational Visual Media*. Springer, 24–41.

Peiying Zhang, Nanxuan Zhao, and Jing Liao. 2025. Style Customization of Text-to-Vector Generation with Image Diffusion Priors. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.

Jin Zhou, Yi Zhou, Hongliang Yang, Pengfei Xu, and Hui Huang. 2026. StrokeFusion: Vector Sketch Generation via Joint Stroke-UDF Encoding and Latent Sequence Diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 13656–13664.

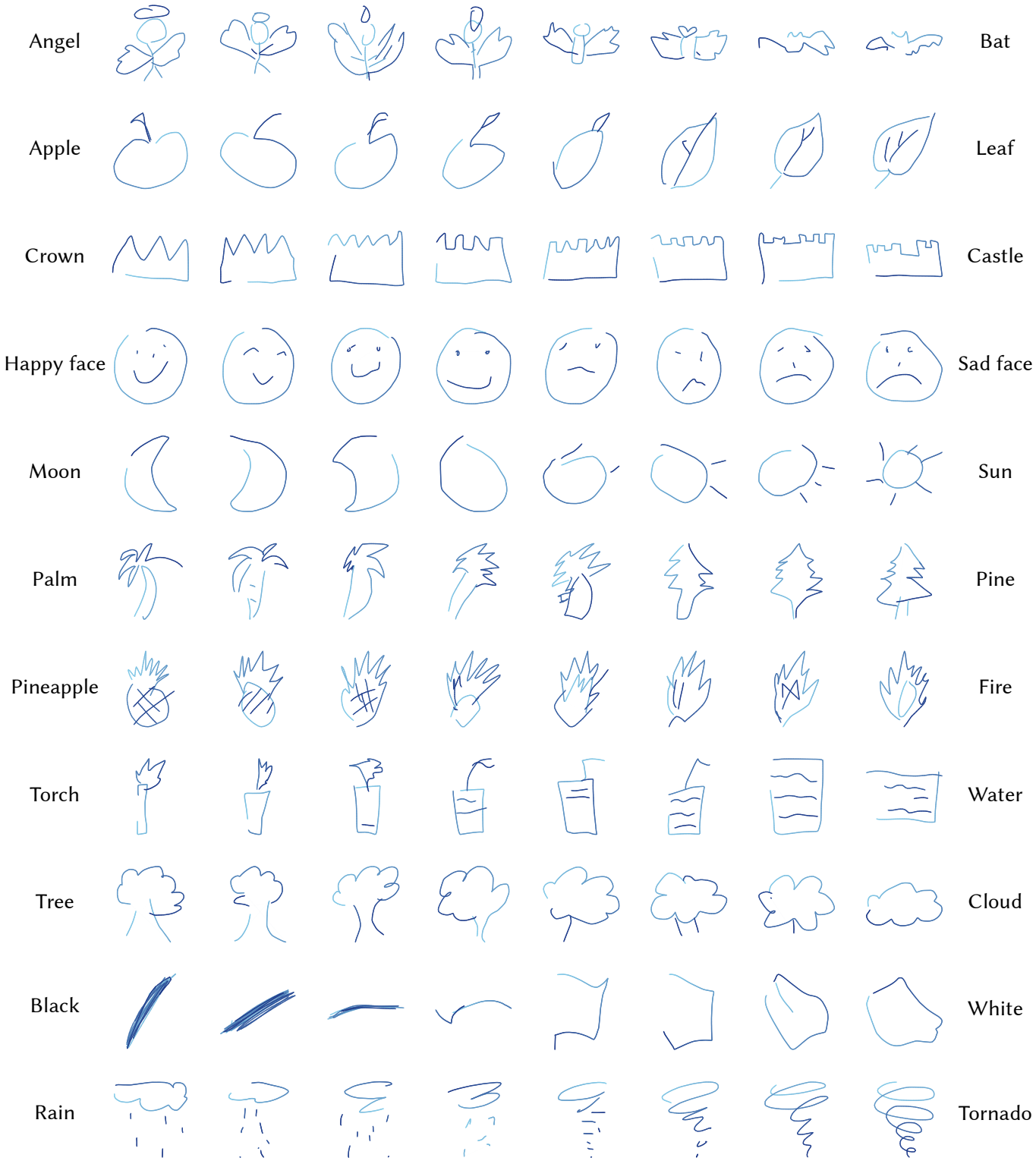

Fig. 9. **Semantic Interpolation in CLIP Space.** Each row visualizes a linear interpolation path between the discrete text embeddings of two distinct concepts (e.g., smoothly morphing from an "Apple" to a "Leaf", or from a "Palm" to a "Pine"). The intermediate generations maintain exceptionally uniform structural transitions and logical stroke topologies without disjointed visual jumps. This progressive morphing provides compelling evidence that our model successfully establishes a continuous generative manifold, rather than stochastically snapping to isolated known modes.

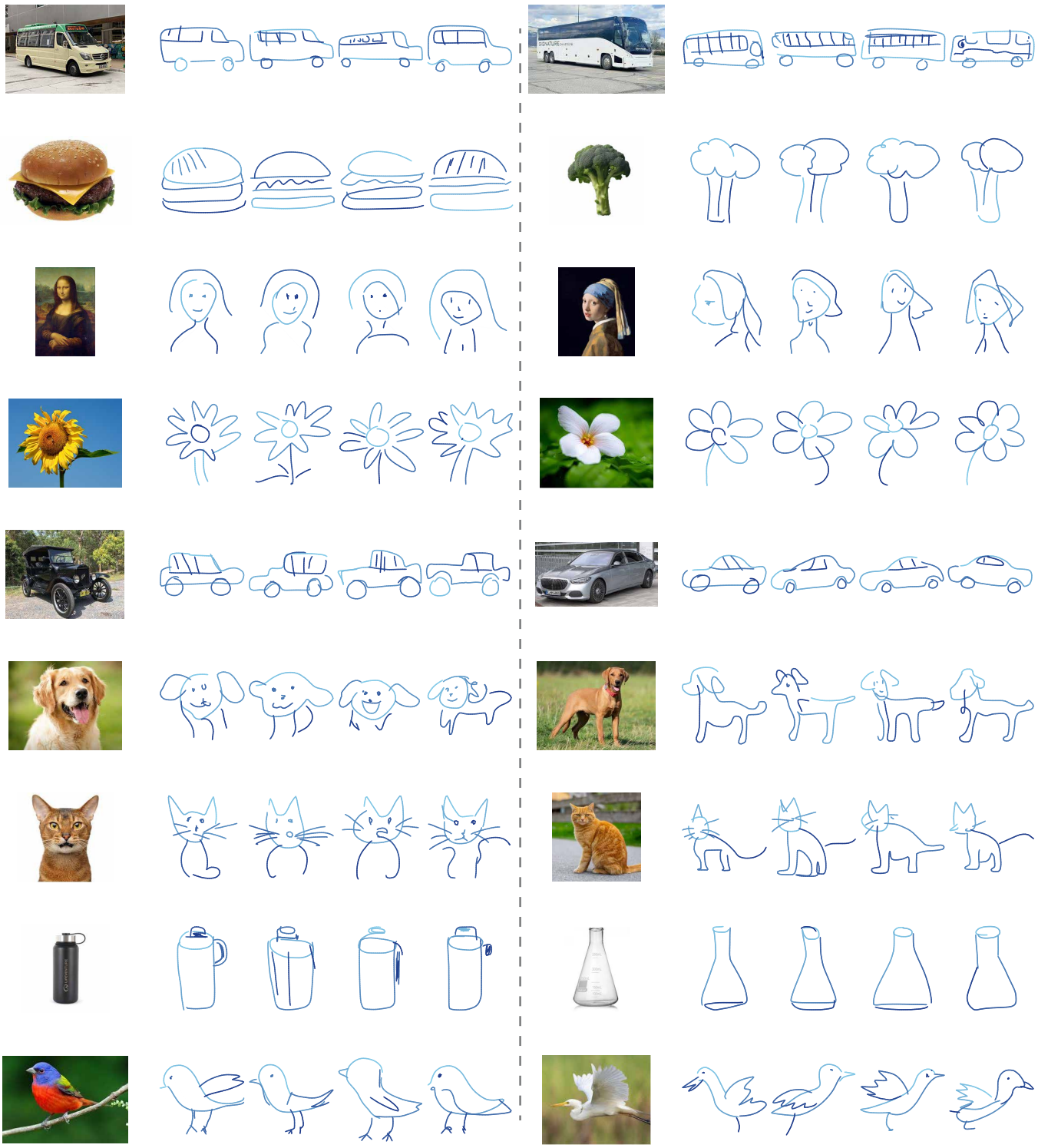

Fig. 10. **Zero-shot Image-Conditioned Vector Sketch Synthesis.** As an intriguing extension, our model can seamlessly accept CLIP visual embeddings in place of textual prompts. This substitution enables a zero-shot stylized redrawing effect, where the generated vector sketches effectively capture the core semantic features and structural characteristics of the reference images.